J Mol Evol (2004) 59:227–238
DOI: 10.1007/s00239-004-2617-8

JOURNAL OF MOLECULAR EVOLUTION

Isochore Structures in the Genome of the Plant *Arabidopsis thaliana*

Ren Zhang,[1] Chun-Ting Zhang[2]

[1] Department of Epidemiology and Biostatistics, Tianjin Cancer Institute and Hospital, Tianjin 300060, China
[2] Department of Physics, Tianjin University, Tianjin, 300072, China

Received: 8 December 2003 / Accepted: 10 February 2004 [*Reviewing Editor*: Martin Kreitman]

Abstract. *Arabidopsis thaliana* is an important model system for the study of plant biology. We have analyzed the complete genome sequences of *Arabidopsis* by using a newly developed windowless method for the GC content computation, the cumulative GC profile. It is shown that the *Arabidopsis* genome is organized into a mosaic structure of isochores. All the centromeric regions are located in GC-rich isochores, called centromere-isochores, which are characterized by a high GC content but low gene and T-DNA insertion densities. This characteristic distinguishes centromere-isochores from the other class of GC-rich isochores, called GC-isochores, which have high gene and T-DNA insertion densities. Consequently, 15 isochores have been identified, i.e., 7 AT-isochores, 3 GC-isochores, and 5 centromere-isochores. The genes in centromere-isochores, which have the highest GC content, have much shorter intron lengths and lower intron numbers, compared to those of the other two types. There is also considerable difference in the numbers and lengths of transposable elements (TEs) between AT and GC-isochores, i.e., the TE number (length) of AT-isochores is 6.3 (7.3) times that of GC-isochores. It is generally believed that TEs are accumulated in the regions surrounding the centromeres. However, within these TE-rich regions, there are regions of extremely low TE numbers (TE deserts), which correspond to the positions of centromere-isochores. In addition, a heterochromatic knob is located at the boundary of an AT-isochore. Furthermore, we show that the differences in GC content among isochores are mainly due to the GC content variation of introns, the third codon positions and intergenic regions.

Key words: Isochores — *Arabidopsis thaliana* — Compositional homogeneity — Windowless technique

Introduction

Arabidopsis thaliana, a flowering plant, is an important model system for the study of plant biology. Based on the sequencing efforts of the *Arabidopsis* Genome Initiative (AGI) established in 1996, the genome sequences of all five chromosomes have been completely sequenced (AGI 2000). The availability of the complete *Arabidopsis* genome sequence provides an unprecedented opportunity to study the global genome organization at the sequence level.

The isochore structure is referred to the phenomenon that in some eukaryotic genomes, the genome is organized into mosaics which are characterized by a fairly constant average GC content over scales of hundreds of kilobases and by abrupt change to another fairly constant-GC-content region (Bernardi 1995; Macaya et al. 1976). More than 10 years ago, Bernardi and coworkers analyzed the isochore structures of plants by density gradient ultracentrifugation experiments (Matassi et al. 1989; Montero et al. 1990; Salinas et al. 1988). The isochore struc-

Correspondence to: Chun-Ting Zhang; *email:* ctzhang@tju.edu.cn

tures of the *Arabidopsis* genome have also been investigated at the sequence level (Nekrutenko and Li 2000; Oliver et al. 2001).

In this report, we analyzed the isochore structures of the *Arabidopsis* genome by a newly developed windowless technique for the GC content computation, the cumulative GC profile. Consequently, 15 isochores have been identified. These isochores have a fairly homogeneous GC content and appear in the genome alternatively, with relatively sharp boundaries. The isochores are classified into three types, i.e., AT-, GC-, and centromere-isochores. The three types are distinct in terms of GC content, gene density and T-DNA insertion density, transposable element (TE) distribution. It is generally believed that TEs are accumulated in the regions surrounding the centromeres. Surprisingly, we found that within these TE-rich regions, there are regions of extremely low TE numbers (TE deserts), which correspond to the position of centromere-isochores. In addition, a heterochromatic knob is located at the boundary of an AT-isochore. The source of GC content variation among isochores was analyzed and shown to be mainly due to the differences of GC content at introns, the third codon positions and intergenic sequences.

Materials and Methods

The genome sequences of *Arabidopsis* were downloaded from http://www.ncbi.nlm.nih.gov. The TE data were based on Wright et al. (2003).

The Z-curve is a three-dimensional space curve constituting the *unique* representation of a given DNA sequence in the sense that each can be *uniquely* reconstructed given the other (Zhang and Zhang 1991, 1994). Based on the Z-curve, any DNA sequence can be uniquely described by three independent distributions, i.e., x_n, y_n, and z_n. In particular, z_n displays the distribution of bases of GC/AT types along the sequence, which is calculated as follows (Zhang and Zhang 1991, 1994)

$$z_n = (A_n + T_n) - (C_n + G_n), \quad n = 0, 1, 2, ..., N, \quad z_n \in [-N, N] \tag{1}$$

where A_n, C_n, G_n, and T_n are the *cumulative* numbers of the bases *A*, *C*, *G*, and *T*, respectively, occurring in the subsequence from the first base to the *n*th base in the DNA sequence inspected, $A_0 = C_0 = G_0 = T_0 = 0$, $z_n = 0$. By viewing the $z_n \sim n$ curve, many global and local features of the GC content can be detected in a perceivable way.

For most genome or chromosome sequences, the curves of $z_n \sim n$ are roughly straight lines. To amplify the variations, the curve of $z_n \sim n$ is fitted by a straight line using the least square technique,

$$z = kn \tag{2}$$

where (z, n) is the coordinate of a point on the fitted straight line and k is its slope. Instead of using the curve of $z_n \sim n$, we will use the $z'_n \sim n$ curve, or simply z′ curve hereafter, where

$$z'_n = z_n - kn \tag{3}$$

Therefore, the variations of $z_n \sim n$ curve deviated from the straight line, which corresponds to a constant GC content (see Eq. [4] below), are protruded by the $z'_n \sim n$ curve. The z′ curve or the cumulative GC profile are used interchangeably in this paper. Let $\overline{GC}$ denote the average GC content within a region Δn in a sequence, it was shown that (Zhang et al. 2001)

$$\overline{GC} = \frac{1}{2}(1 - k - \frac{\Delta z'_n}{\Delta n}) \equiv \frac{1}{2}(1 - k - k') \tag{4}$$

where $k' = \Delta z'_n/\Delta n$ is the average slope of the z′ curve within the region Δn. It is clear to see from Eq. (4) that an up jump in the z′ curve, i.e., $k' > 0$, indicates a decrease in GC content or an increase in AT content, whereas a drop in the z′ curve, i.e., $k' < 0$, indicates an increase in GC content or a decrease in AT content. In addition, if a region in the z′ curve is a purely (approximately) straight line, then the GC content keeps absolutely (approximately) constant within this region. Any sharp maximum (minimum) point in the z′ curve indicates a turning point, where the GC content undergoes an abrupt change from a relatively GC-poor (GC-rich) region to a relatively GC-rich (GC-poor) region. The region Δn is usually chosen to be a fragment of a natural DNA sequence, e.g., isochore. The above method to calculate the GC content is called the windowless technique (Zhang et al. 2001).

The concept of isochores is related to domains of relatively homogeneous GC content with large scales in genomes, in which the variations of GC content may be considered to be small. Based on the z′ curve, a homogeneity index h, which describes the smallness of the GC content variations in isochores, was introduced (Zhang and Zhang 2003):

$$h = d_{\text{isochore}}/d_{\text{chromosome}} \tag{5}$$

where

$$d_{\text{isochore}} = \sqrt{\sum_{n=1}^{M} (z'_n)^2/M} \tag{6}$$

$$d_{\text{chromosome}} = \sqrt{\sum_{n=1}^{N} (z'_n)^2/N} \tag{7}$$

where $z'_n \sim n$ is the cumulative GC profile defined in Eq. (3) for the isochore and the entire chromosome studied, respectively, and M and N are their lengths. If $h \ll 1$, the variations of the GC content of the isochore may be considered to be small. No *prior* knowledge is available to define isochores based on h. In the current study, we arbitrarily chose $h = 0.2$ as the threshold of isochores.

One-way ANOVA tests were performed when comparing multiple groups of samples, whereas Student *t*-tests were performed when comparing two groups of samples, unless indicated otherwise.

Results

Features of the z′ Curves for the Five Chromosomes of Arabidopsis

The z′ curves for all five chromosomes of *Arabidopsis* are shown in Fig. 1. The cumulative GC profiles show that in these chromosomes long domains that have relatively homogeneous GC content exist, as reflected by the fact that some regions of the cumulative GC profiles can be approximately described by straight lines. The domains that have relatively homogeneous GC content are isochores. A quantitative index is used to assess the homogeneity of the GC content of isochores and this is discussed in an-

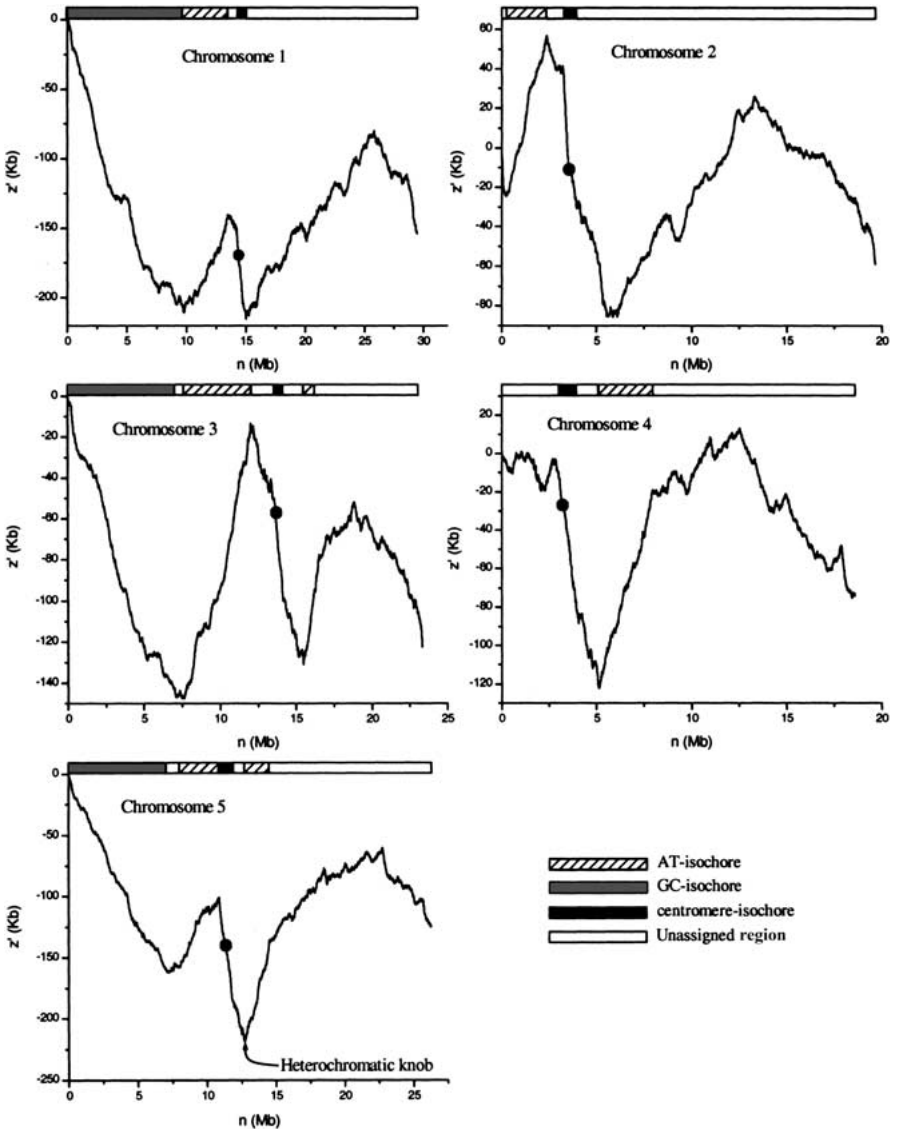

Fig. 1. The cumulative GC profiles for the five chromosomes of the *Arabidopsis* genome. The filled circles indicate the position of centromeric regions. The position of the heterochromatic knob is indicated by an arrow. An up jump in the curve indicates a decrease in GC content, whereas a drop in the curve indicates an increase in GC content. If a region is approximately described by a straight line, then the GC content is approximately constant, suggesting this region to be an isochore. Therefore, the sharp peaks of the curves suggest a mosaic structure of the genome, i.e., the GC content undergoes abrupt changes, from GC-rich regions to GC-poor regions, alternatively, and vice versa. In addition, the variations in GC content in isochores are relatively small. Refer to the text for a quantitative definition. All the centromeric regions are located in GC-rich isochores. The overall patterns GC distributions of chromosomes 1, 3, and 5 are similar. The locations of isochors and unassigned regions are indicated by pars of different patterns.

other section. In addition, the GC-rich isochores are followed immediately by AT-rich isochores, and vice versa, clearly indicating a mosaic structure of the genome. In addition, one striking feature is that all the centromeric regions of the five chromosomes are located within GC-rich isochores. Furthermore, the overall patterns of GC content variation can be roughly classified into two groups: those of chromosomes 1, 3, and 5 are highly similar, and those of chromosomes 2 and 4 are highly similar.

Isochores of Arabidopsis *and Their Classification*

A total of 15 isochores have been identified in the *Arabidopsis* genome. In a previous work (Zhang and Zhang 2003), we classified the isochores into two types, i.e., GC-isochores and AT-isochores. The GC (AT-)-isochores are isochores whose GC content is higher (lower) than that of the chromosome where the isochores are located. The gene density of GC-isochores is usually high, whereas that of AT-isochores is usually low. In the *Arabidopsis* genome, all the centromeric regions are located in five GC-rich isochores. Although these five GC-rich isochores have a high GC content, they are distinct from the other class of GC-rich isochores in terms of many features, such as gene and T-DNA insertion densities. A much lower gene density was found in these five GC-isochores that are associated with centromeric regions (Table 1). Therefore, we classify the isochores that are associated with the centromeric regions as another class, the centromere-isochores.

Gene Distribution Among Isochores

In mammalian genomes, genes are preferentially distributed in high GC regions (Bernardi 1995;

Table 1. The isochores in the *Arabidopsis* genome[a]

No. isochores	No. chromo-somes	Type	Start (Mb)	End (Mb)	Length (Mb)	Gene No. per Mb	T-DNA No. per Mb	TE No./Mb	TE length/Mb	Intron No./gene	Intron length
1	1	GC	0.00	9.74	9.74	280.70	3529.47	16.94	23,665.61	4.55 ± 5.12	730.35 ± 852.50
2	1	AT	9.74	13.48	3.74	229.41	2695.19	58.02	88,182.89	4.12 ± 4.66	713.96 ± 891.55
3	1	Centromere	14.15	14.90	0.75	110.67	1720.12	12.00	25,910.67	4.18 ± 4.02	673.81 ± 630.56
4	2	AT	0.23	2.36	2.13	230.52	2403.76	104.23	13,8371.8	3.56 ± 4.03	588.84 ± 690.31
5	2	Centromere	3.24	3.60	0.36	163.89	1658.33	52.78	48,883.33	1.67 ± 4.02	292.92 ± 667.98
6	3	GC	0.00	7.11	7.11	292.69	3571.87	9.99	7,523.629	4.54 ± 5.32	703.18 ± 838.50
7	3	AT	7.65	11.95	4.3	237.91	2520.70	62.56	61,615.58	3.36 ± 4.00	552.72 ± 692.01
8	3	Centromere	13.56	14.05	0.49	18.37	1430.61	34.69	74,291.84	1.67 ± 1.23	245.78 ± 189.56
9	3	AT	15.50	16.45	0.95	189.47	2174.74	120.00	167,442.1	3.56 ± 3.94	626.26 ± 715.84
10	4	Centromere	3.04	3.90	0.86	29.07	1704.65	10.47	18,605.81	3.28 ± 2.84	564.56 ± 747.90
11	4	AT	5.12	7.88	2.76	224.64	2610.14	53.99	49,235.14	3.72 ± 3.94	684.07 ± 771.49
12	5	GC	0.00	7.17	7.17	300.28	3494.98	13.25	20,119.94	4.66 ± 5.09	750.04 ± 830.86
13	5	AT	8.02	10.20	2.18	211.01	2443.58	75.23	113,392.7	4.59 ± 5.40	724.81 ± 906.26
14	5	Centromere	11.05	11.89	0.84	61.90	1509.52	15.48	6,617.857	2.70 ± 2.11	476.58 ± 434.80
15	5	AT	12.73	14.12	1.39	180.58	1930.46	115.11	253,079.1	3.75 ± 4.11	622.96 ± 810.91
	1	Unassigned	14.90	30.33	15.43	251.85	2943.94	32.08	39,163.38	4.42 ± 5.10	720.70 ± 899 15
	2	Unassigned	3.60	19.64	16.04	254.80	3002.24	43.83	69,527.99	4.12 ± 4.85	696.73 ± 844.46
	3	Unassigned	16.45	23.33	6.88	283.28	3483.72	30.96	34,804.36	4.16 ± 4.74	647.13 ± 800.69
	4	Unassigned	7.88	18.58	10.7	269.72	2957.85	12.43	12.534.49	4.42 ± 4.56	729.74 ± 817.90
	5	Unassigned	14.12	26.26	12.14	278.09	3343.82	36.16	48,730.97	4.14 ± 4.77	635.86 ± 745.95

			GC, content	GC, exon	GC, intron	GC, intergenic	GC12	GC3	*h*	*k*
1	1	GC	0.364	0.441 ± 0.047	0.335 ± 0.043	0.330 ± 0.033	0.452 ± 0.038	0.437 ± 0.063	0.19	0.29
2	1	AT	0.344	0.436 ± 0.046	0.330 ± 0.050	0.320 ± 0.035	0.452 ± 0.041	0.424 ± 0.064	0.03	0.29
3	1	Centromere	0.395	0.451 ± 0.055	0.350 ± 0.068	0.321 ± 0.117	0.448 ± 0.054	0.460 ± 0.072	0.04	0.29
4	2	AT	0.338	0.430 ± 0.048	0.329 ± 0.054	0.312 ± 0.036	0.445 ± 0.043	0.408 ± 0.060	0.08	0.29
5	2	Centromere	0.425	0.449 ± 0.048	0.379 ± 0.073	0.423 ± 0.046	0.457 ± 0.038	0.427 ± 0.068	0.10	0.29
6	3	GC	0.371	0.443 ± 0.046	0.336 ± 0.042	0.331 ± 0.034	0.452 ± 0.039	0.439 ± 0.063	0.14	0.28
7	3	AT	0.346	0.434 ± 0.043	0.327 ± 0.050	0.315 ± 0.036	0.447 ± 0.044	0.418 ± 0.061	0.09	0.28
8	3	Centromere	0.403	0.507 ± 0.049	0.471 ± 0.080	0.408 ± 0.020	0.483 ± 0.043	0.517 ± 0.048	0.01	0.28
9	3	AT	0.336	0.432 ± 0.049	0.333 ± 0.060	0.315 ± 0.045	0.443 ± 0.043	0.428 ± 0.065	0.12	0.28
10	4	Centromere	0.400	0.464 ± 0.067	0.406 ± 0.083	0.398 ± 0.052	0.452 ± 0.041	0.468 ± 0.085	0.05	0.28
11	4	AT	0.334	0.432 ± 0.046	0.331 ± 0.050	0.316 ± 0.039	0.449 ± 0.039	0.417 ± 0.063	0.06	0.28
12	5	GC	0.368	0.442 ± 0.046	0.336 ± 0.042	0.329 ± 0.033	0.454 ± 0.037	0.436 ± 0.062	0.06	0.29
13	5	AT	0.347	0.437 ± 0.048	0.327 ± 0.051	0.316 ± 0.036	0.450 ± 0.036	0.429 ± 0.063	0.03	0.29
14	5	Centromere	0.400	0.467 ± 0.067	0.387 ± 0.087	0.393 ± 0.038	0.469 ± 0.040	0.468 ± 0.071	0.04	0.29
15	5	AT	0.336	0.436 ± 0.052	0.346 ± 0.064	0.328 ± 0.058	0.442 ± 0.046	0.406 ± 0.072	0.16	0.29
	1	Unassigned	0.351	0.436 ± 0.046	0.331 ± 0.046	0.322 ± 0.035	0.450 ± 0.038	0.422 ± 0.060	0.23	0.29
	2	Unassigned	0.359	0.438 ± 0.048	0.334 ± 0.048	0.324 ± 0.035	0.453 ± 0.040	0.427 ± 0.064	1.37	0.29
	3	Unassigned	0.364	0.439 ± 0.045	0.335 ± 0.045	0.327 ± 0.033	0.452 ± 0.036	0.432 ± 0.061	0.40	0.28
	4	Unassigned	0.365	0.441 ± 0.047	0.335 ± 0.047	0.327 ± 0.035	0.455 ± 0.039	0.433 ± 0.063	0.72	0.28
	5	Unassigned	0.356	0.436 ± 0.046	0.330 ± 0.044	0.319 ± 0.034	0.450 ± 0.040	0.422 ± 0.061	0.61	0.29

[a] The GC content is calculated using Eq. (4). The homogeneity index *h* of the GC content of isochores is defined in Eq. (5). The slope *k* is defined using Eq. (2).

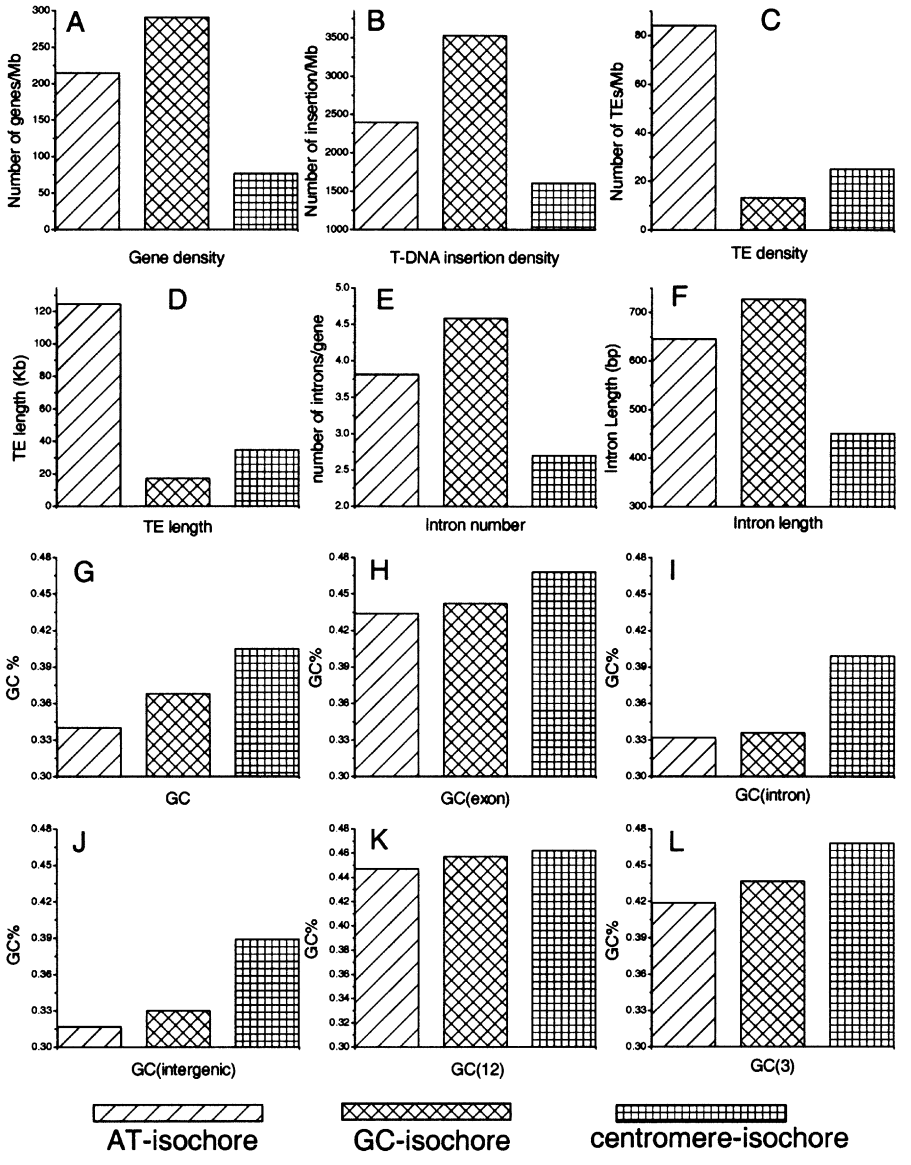

Fig. 2. Different genome features among three classes of isochores. These features include (**A**) gene density, (**B**) T-DNA insertion density, (**C**) TE number, (**D**) TE length, (**E**) intron number, (**F**) intron length, (**G**) GC, (**H**) GC (exon), (**I**) GC (intron), (**J**) GC (intergenic), (**K**) GC12, and (**L**) GC3.

Lander et al. 2001). In human isochores, a higher gene density was found in H3 isochores (more GC-rich) than in L isochores (less GC-rich) (Bernardi 2000), which was subsequently confirmed based on isochores identified by a compositional segmentation method (Oliver et al. 2002). This correlation also holds, generally, for the *Arabidopsis* genome (Fig. 2A). The average GC contents for AT-isochores and GC-isochores are 0.340 and 0.368, respectively (Table 2) ($p < 0.01$). The gene density in AT-isochores and GC-isochores is 215 and 291/Mb, respectively ($p < 0.001$). The third type of isochores, centromere-isochores, although they have the highest GC content, 0.405 ($p < 0.0001$ compared with that of AT-isochores and $p < 0.01$ compared with that of GC-isochores), have the lowest gene density, 77/Mb ($p < 0.001$ compared with that of GC-isochores). Therefore, the centromere-isochores are distinct from the other class of GC-rich isochores, GC-isochores, which are characterized by a high GC content and a high gene density.

T-DNA Insertion Site Distribution Among Isochores

The distribution of the integration of transferred DNA (T-DNA) was investigated in the genome of *Arabidopsis* in 2000 (Barakat et al. 2000). Recently, over 225,000 independent *Agrobacterium* T-DNA insertion events in the *Arabidopsis* genome have been created that represent near-saturation of the gene space. The precise locations were determined for more than 88,000 T-DNA insertions. Genome-wide analysis of the distribution of integration events revealed the existence of a large integration site bias at the chromosome level (Alonso et al. 2003).

We studied the distribution of T-DNA insertion sites among isochores and found that T-DNA is most preferentially integrated into GC-isochores and most

Table 2. Statistics of the three types of isochores in the *Arabidopsis* genome[a]

		Gene No./Mb		T-DNA insertion No./Mb		TE No./Mb	
Type	N	Average	SD	Average	SD	Average	SD
AT	7	214.79	22.05	2396.94	264.24	84.16	28.24
GC	3	291.22	9.87	3532.11	38.51	13.39	3.48
Centromere	5	76.78	60.51	1604.65	128.02	25.08	18.28
Unassigned	5	267.55	13.90	3146.31	250.04	31.09	11.59
		TE length/Mb		Intron No./gene		Intron length	
		Average	SD	Average	SD	Average	SD
AT		124474.19	70273.97	3.81	0.42	644.80	64.75
GC		17103.06	8483.34	4.58	0.07	727.86	23.53
Centromere		34861.90	26895.31	2.70	1.08	450.73	180.48
Unassigned		40952.24	20770.45	4.25	0.15	686.03	42.59
		GC		GC exon		GC intron	
		Average	SD	Average	SD	Average	SD
AT		0.340	0.01	0.434	0.00	0.332	0.00
GC		0.368	0.00	0.442	0.00	0.336	0.00
Centromere		0.405	0.01	0.468	0.02	0.399	0.05
Unassigned		0.359	0.01	0.438	0.00	0.333	0.00
		GC, intergenic		GC12		GC3	
		Average	SD	Average	SD	Average	SD
AT		0.317	0.01	0.447	0.00	0.419	0.01
GC		0.330	0.00	0.453	0.00	0.437	0.00
Centromere		0.389	0.04	0.462	0.01	0.468	0.03
Unassigned		0.324	0.00	0.452	0.00	0.427	0.01

Note. One-way ANOVA tests were performed, which showed that all the listed features among three types of isochores are statistically different, at $p < 0.01$.

unfavorably integrated into centromere-isochores. The T-DNA insertion densities for AT, GC, and centromere-isochores are 2397, 3532, and 1605 sites/Mb, respectively ($p < 0.0001$) (Table 2 and Fig. 2B).

Although the precise mechanism of T-DNA integration in the host genome is not fully understood, one possible reason for the biased integration sites is that the biased integration is due to the different chromatin structures. For example, the integration may be promoted by increased chromatin accessibility in transcribed regions, thereby removing inhibitory effects of unfavorable chromatin environment (Schroder et al. 2002). Indeed, it has recently been found that sites of HIV integration in the human genome are not randomly distributed but instead are enriched in active genes (Schroder et al. 2002). Therefore, the biased distribution of T-DNA insertion sites among these isochores may reflect the difference in the chromatin structures of the three classes of isochores.

Transposable Element Distribution Among Isochores

Transposable elements (TEs) have been found in all eukaryotes, and these elements have the ability to move into new locations in chromosomes. The locations of TEs are highly biased, and in the *Arabidopsis* genome, it is generally believed that TEs are accumulated in the regions surrounding the centromeres (Copenhaver et al. 1999; AGI 2000; Wright et al. 2003). Recently, the locations of all TEs in the *Arabidopsis* genome were determined based on an *Arabidopsis* TE database (http://www.tebureau.mcgill.ca/). By using these data, we analyzed the TE distribution among isochores.

Based on the study of TE distribution along genomes, it is shown that regions surrounding the centromeric regions indeed contain high concentration of TEs. However, we noticed that in these TE-rich regions, there are regions that have extremely low numbers of TEs (TE deserts). Strikingly, these TE deserts correspond to the locations of centromere-isochores. Refer to Fig. 3 for an example based on chromosome 1. One possible explanation is that the regions corresponding to TE deserts are critical for centromere functions, and therefore, insertions of TEs into these regions are deleterious and eliminated by natural selection.

There is also considerable difference of the TE numbers and lengths between the AT- and the GC-isochores. The average TE numbers of AT- and GC-isochores are 84.16 and 13.39/Mb, respectively; the

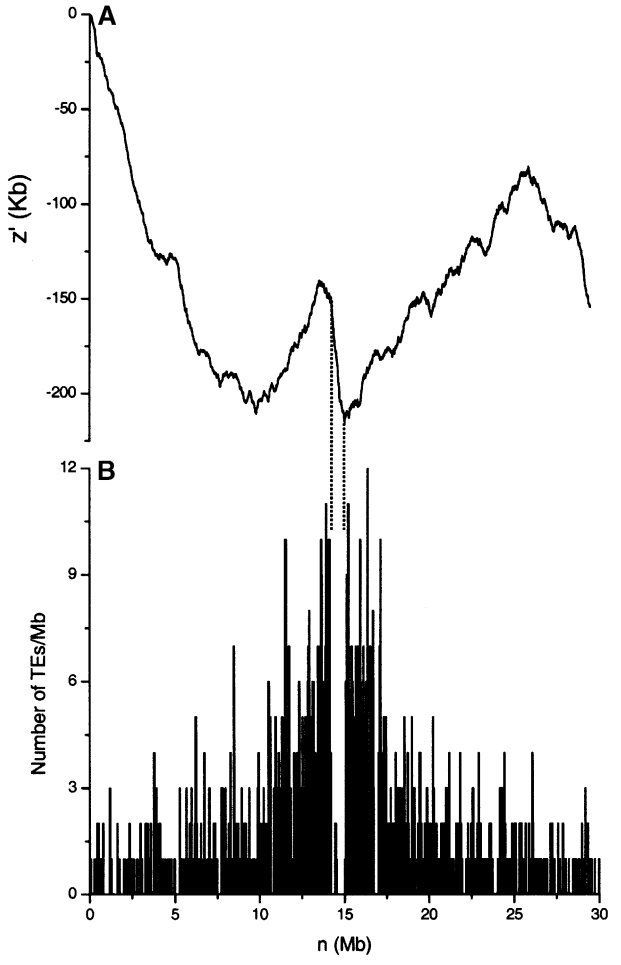

Fig. 3. **A** The z' curve for chromosome 1. **B** TE distribution of chromosome 1 based on 50-kb sliding windows. Note that TEs are accumulated in the region surrounding the centromeres. However, in the IE-rich region, there is a segment of genome sequence that has an extremely low number of TEs (TE desert). The TE desert corresponds to the position of centromere-isochores.

average TE lengths are 124 and 17 kb/Mb, respectively (Table 2 and Figs. 2C and D) ($p < 0.01$). Therefore, although there is only a 2.8% difference in GC content, the TE number (length) of AT-isochores is 6.3 (7.3) times that of GC-isochores. As mentioned previously, AT-isochores have a lower gene density than GC-isochores. The distributions of TEs in AT- and GC-isochores are consistent with the result of Wright et al. (2003), which shows a negative correlation between gene density and TE abundance.

Intron Number and Length Distributions Among Isochores

It has been found that two classes of genes are present in *Arabidopsis*. One class of genes, the GC-rich class, has relatively low intron numbers and short concatenated intron lengths. The other class of genes, the GC-poor class, has relatively high intron numbers and long concatenated intron lengths (Barakat et al. 1998; Carels and Bernardi 2000). We analyzed the intron length and number distributions among isochores. The intron numbers of AT, GC, and centromere-isochores are 3.81, 4.58, and 2.70, respectively ($p < 0.001$); intron lengths are 645, 728, and 450 bp, respectively ($p < 0.001$) (Table 2). Therefore, genes within the most GC-rich class of isochores, centromere-isochores, have a much lower intron number and shorter intron length, than those of the other two classes (Figs. 2D and E). However, genes in the GC-isochores do not have lower intron numbers and shorter intron lengths. Therefore, it is likely that in the previously found two classes of genes (Carels and Bernardi 2000), the GC-rich class, which has low intron numbers and short lengths, contains the genes in the centromere-isochores.

A Heterochromatic Knob Is Located at an Isochore Boundary

The heterochromatic knobs were first observed by McClintock (1929) in the maize genome. These knobs are cytologically detectable, darkly stained heterochromatic regions present on the maize pathytene chromosomes. Many genetic effects are linked to the heterochromatic knobs. For instance, in the maize genome, the heterochromatic knobs were found to affect the recombination frequency and chromosome behavior in microspore divisions (Rhoades 1978; Rhoades and Dempsey 1973). Chromosome 5 of the *Arabidopsis* genome has a heterochromatic knob (AGI 2000) and the location is at the boundary of an AT-isochore. At the location of the heterochromatic knob, the genome undergoes a relatively abrupt change from a GC-rich region to an AT-rich region. However, chromosome 4 also has a heterochromatic knob, which is not close to any isochore boundary. Therefore, there is a possibility that the correlation between the heterochromatic knob and the isochore boundary in chromosome 5 is due to coincidence. Further information on heterochromatic knobs is needed to investigate this issue. If there is indeed a correlation between heterochromatic knobs and isochores, the correlation between these two structures may provide further insight into the origin of isochores. In addition, the cumulative GC profiles for chromosomes 1, 3, and 5 show a similar overall pattern (Fig. 1), therefore, it will be interesting to investigate the corresponding parts of chromosomes 1 and 3, to examine whether there are also heterochromatic knob structures.

Features of Unassigned Regions

In each chromosome, besides isochore regions, there are other unassigned regions, which are not iso-

chores. We have also studied various features of these unassigned regions (Tables 1 and 2). The average gene density of these regions is 267.55/Mb; the T-DNA insertion number is 3146.31/Mb; the TE number is 31.09/Mb; the TE length is 40,952.24/Mb; the intron number is 4.25/gene; the intron length is 686.03/gene; and the GC contents of exons, introns, intergenic regions, GC12, and GC3 are 0.438, 0.333, 0.324, 0.452, and 0.427, respectively. In brief, all these features are in between those of AT- and GC-isochores. Because the GC content of these regions is in between those of the AT- and GC-isochore, it appears that the GC content is a critical factor in determining all these features.

Discussion

Source of the GC Content Variations Among Isochores

To investigate the source of the GC content variation among isochores, we calculated the GC content of exons, introns, GC12, GC3, and intergenic regions among different classes of isochores (Tables 1 and 2 and Fig. 2). Generally, the GC contents of all these different regions of AT-isochores are less than those of GC-isochores, which are less than those of centromere-isochores. For the AT-, GC-, and centromere-isochores, the GC contents of exons are 0.434, 0.442, and 0.468, respectively ($p < 0.01$); the GC contents of introns are 0.332, 0.336, and 0.399, respectively ($p < 0.01$); the GC contents of intergenic regions are 0.317, 0.330, 0.389, respectively ($p < 0.01$); the GC contents of GC12 are 0.447, 0.453, and 0.462, respectively ($p < 0.05$); and the GC contents of GC3 are 0.419, 0.437, and 0.468, respectively ($p < 0.01$). There is a clear difference, however, between GC12 and the GC content of noncoding regions, i.e., the GC content differences in noncoding regions among isochores are much more than those of GC12. For instance, the difference in GC12 between centromere and GC-isochore is 0.009, whereas the difference in GC content of introns is 0.063; the difference in GC3 is 0.031; and the difference in GC content of intergenic regions is 0.059. Therefore, the difference in isochore GC content is likely to be largely due to the variation in GC content in noncoding regions. This observation is consistent with the GC variation in human isochores, i.e., GC3 variation is greater than the GC content variation of isochores (Clay et al. 1996). These noncoding regions are believed to have less selective pressure than coding regions. Therefore, this GC variation pattern appears to support the view that isochores are due to the mutational bias along genomes (Eyre-Walker 1999; Eyre-Walker and Hurst 2001).

Comparison of Arabidopsis *Isochores with Those of the Human Genome*

Recently, the isochore structures of the human genome have been identified based on the cumulative GC profile (Zhang and Zhang 2003). The isochore structures of *Arabidopsis* are distinct from those of human in several aspects. First, the variations in GC content between AT- and GC-isochores are different. In the human genome, the average GC contents for AT- and GC-isochores are 0.38 and 0.47, respectively. Therefore, GC-isochores are 9% higher in GC content than AT-isochores. In the *Arabidopsis* genome, the average GC contents for AT- and GC-isochores are 0.34 and 0.37, respectively. Therefore, there is only a 3% difference in terms of GC content between these two types of isochores. However, isochore structures are still clear even though the GC content difference is not as large as those of the human genome. Another striking difference is that in the human genome, the type centromere-isochore does not exist. The behaviors of the GC-isochores that harbor the centromeric regions show no difference from that of other GC-isochores in the human genome. This difference may reflect the characteristic organization of the centromeric region of *Arabidopsis* (Haupt et al. 2001). In addition, the relative proportion of the AT-, GC-, and centromere-isochores in the *Arabidopsis* genome is 17.45, 53.65, and 7.37%, respectively. However, in the human genome, the GC-rich isochores (H3) only represent ~3–5% (Bernardi 1995). Furthermore, in the *Arabidopsis* genome, the GC-isochores (8.01 Mb) are longer than the AT-isochores (2.49 Mb) and there are fewer GC-isochores (3) than AT-isochores (7), whereas in the human genome, GC-isochores (11.25 Mb) are shorter than AT-isochores (13.49 Mb) and there are more GC-isochores (34) than AT-isochores (22).

Comparison of the Cumulative GC Profile with the GC Content Distribution Based on Other Methods

As a routine procedure, the GC content distribution is computed by a sliding-window technique (Lander et al. 2001; Waterston et al. 2002). In this method, the number of G and C residues are counted within a window, therefore, the size of the window can be considered as the resolution of the GC content. The resolution of this method is usually low, because a small window size leads to large statistical fluctuations. In addition, the GC content distribution obtained by window-based methods is dependent on the window sizes chosen; in contrast, the cumulative GC profile is unique for a genome. A comparison between the window-based and the windowless approaches has been detailed in the literature (Li 2001;

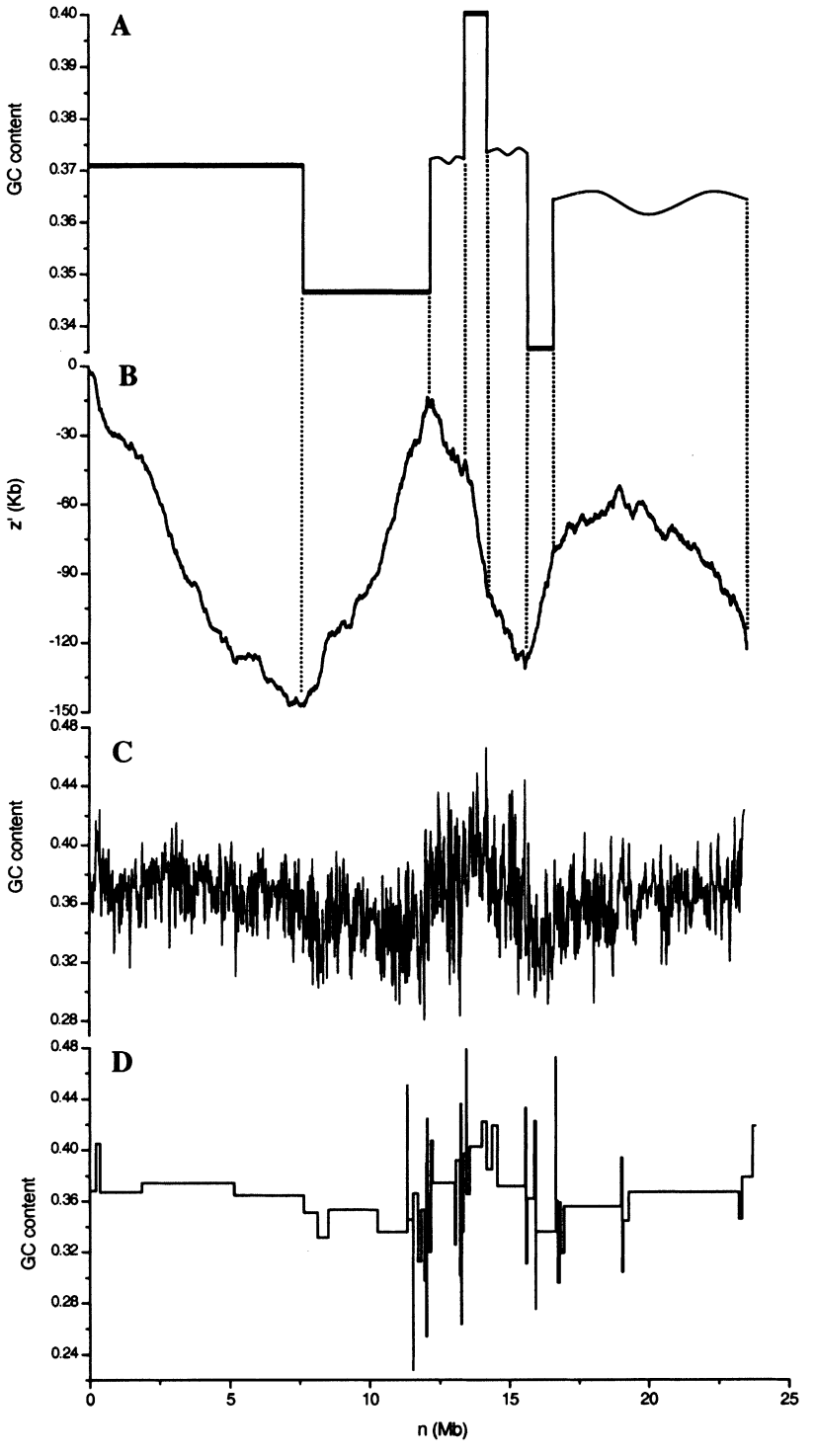

Fig. 4. **A** Schematic diagram showing the GC content of isochores in *Arabidopsis thaliana* chromosome 3 based on the cumulative GC profile. The regions of isochores are marked with bold horizontal lines, whereas those of interisochores are marked with wavy lines. **B** The z' curve for the chromosome 3. **C** The GC content calculated based on the 20-kb sliding window technique. Note that the boundaries of isochores cannot be identified based on the window technique. **D** Fifty-six isochores indentified by the entropic segmentation method.

Zhang and Zhang 2003). We want to emphasize that in genomes with larger GC content variation, such as the human genome, the homogeneous GC content domains can be revealed by window-based methods, although the boundaries sometimes cannot be determined precisely (Pavlicek et al. 2002). However, for the genomes with relative low GC content variations, the disadvantages of window-based methods are more obvious, i.e., the window-based methods usually show a complex pattern. For example, the GC content distribution of the *Arabidopsis* genome is plotted based on 20-kb sliding windows (Fig. 4C), which shows a complex pattern, and boundaries between domains of the GC distribution are totally blurred by the variations. On the contrary, the z' curve clearly shows five domains of GC distribution, i.e., two AT-isochores, two GC-isochores, and a domain with large GC variation (Figs. 3A and B). Therefore,

the isochore structures and their boundaries can be revealed by the cumulative GC profiles clearly.

Another windowless tool to analyze genome heterogeneity is compositional segmentation, and this technique has also been used to study the isochore structures of eukaryotic genomes (Li 2001; Oliver et al. 2001). The isochores of the *Arabidopsis* genome has also been studied by the entropic segmentation method (Oliver et al. 2001). Please also visit http://bioinfo2.ugr.es/isochores/ for details about the method. As a comparison, the isochores of chromosome 3, obtained based on the entropic segmentaton method, are shown in Fig. 4D. One apparent difference is that much more isochores (56) were determined by the entropic segmentation method than those by the z' curve. Another difference is that all regions of the chromosome are classified into different isochores based on the entropic segmentation method, whereas there are some unassigned regions, which are not isochores, based on the z' curve. However, some segmentation points of both methods are highly consistent. For instance, the isochore boundaries obtained based on the z' curve overlap well with some of the segmentation points based on the entropic segmentation method. In addition, among the 56 isochores obtained based on the entropic segmentation method, many boundaries correspond to some jumps in the z' curve. Therefore, in some aspects, the two methods are consistent. However, the z' curve appears to be more intuitive and can give a picture of the global GC content distribution along genomes.

Definition of Isochores

Identification of isochore structures in eukaryotic genomes provides much insight into the understanding of the genome organization, because of the clear functional implications of isochores. For instance, isochores have been correlated with gene density (Zoubak et al. 1996), chromosome bands (Saccone et al. 1993), and repeat elements (Meunier-Rotival et al. 1982). In the *Arabidopsis* genome, the isochores identified have been shown to be related to gene density, T-DNA insertion density, TE density, intron length, and so on. Although isochores have been known for more than 25 years, currently no clear definition of isochores is available. We defined the index, h, to assess the relative homogeneity of isochores compared to the variation in GC content of the whole genome. In this study, we arbitrarily chose a threshold of the homogeneity index, h, to be 0.20 for isochores. In fact, the homogeneity index, h, is more suitable to be an index to assess the relative homogeneity of isochores, rather than a definition. By using the entropic segmentation method, genomes can be split into many segments (isochores) objectively, based on segmentation points (Oliver et al. 2001). For instance, 56 isochors were found in chromosome 3 of the *Arabidopsis* genome (Oliver et al. 2001). However, it seems to be unreasonable that every region of a genome is an isochore. Therefore, due to the lack of a clear definition, most isochores identified so far are quite subjective.

The homogeneity of the GC content of isochores should be considered to be *relative*, whereas boundaries of isochores are *absolute*. No strict isochores that have an absolutely constant GC content have been found in the human genome, as well as other genomes. In terms of the homogeneity index h, h cannot be equal to 0. In some sense, the homogeneity of GC content of isochores is not as important as their functional implications. Isochores are a segment of genome DNA sequences, in which many characteristics, such as the gene density and repeat density, are different from those of other isochores (Meunier-Rotival et al. 1982; Zoubak et al. 1996). Therefore, isochores may be deemed as function domains of genomes or chromosomes, whose boundaries have critical biological meanings. For example, the boundary between Class II and Class III isochores in the human MHC sequence correspond to the change in replication timing (Tenzen et al. 1997). In the *Arabidopsis* genome, the centromere-isochores correspond to TE deserts. The problem is how to find these boundaries both experimentally and theoretically. The cumulative GC profile is one of the available tools (Li et al. 2002; Oliver et al. 2001; Peshkin and Gelfand 1999) to determine the isochore boundaries. The characterization of isochores and their boundaries will provide a solution to define isochores based on their biological functions, and in this regard, the isochore boundary appears to be more important than the homogeneity of GC content of isochores.

Acknowledgments. The present study was supported in part by the 973 Project of China (Grant 1999075606). We are indebted to Dr. Joseph Ecker, who kindly provided us the T-DNA insertion data used in the present study. We cordially thank the referees for their comments and suggestions, which were critical in improving the quality of the current article. We also thank Feng Gao, who helped prepare the data for GC12 and GC3.

References

Alonso JM, Stepanova AN, Leisse TJ, Kim CJ, Chen H, Shinn P, Stevenson DK, Zimmerman J, Barajas P, Cheuk R, Gadrinab C, Heller C, Jeske A, Koesema E, Meyers CC, Parker H, Prednis L, Ansari Y, Choy N, Deen H, Geralt M, Hazari N, Hom E, Karnes M, Mulholland C, Ndubaku R, Schmidt I, Guzman P, Aguilar-Henonin L, Schmid M, Weigel D, Carter DE, Marchand T, Risseeuw E, Brogden D, Zeko A, Crosby WL, Berry CC, Ecker JR (2003) Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. Science 301:653–657

Barakat A, Matassi G, Bernard G (1998) Distribution of genes in the genome of *Arabidopsis thaliana* and its implications for the genome organization of plants. Proc Natl Acad Sci USA 95:10044–10049

Barakat A, Gallois P, Raynal M, Mestre-Ortega D, Sallaud C, Guiderdoni E, Delseny M, Bernardi G (2000) The distribution of T-DNA in the genomes of transgenic *Arabidopsis* and rice. FEBS Lett 471:161–164

Bernardi G (1995) The human genome: Organization and evolutionary history. Annu Rev Genet 29:445–476

Bernardi G (2000) Isochores and the evolutionary genomics of vertebrates. Gene 241:3–l7

Carels N, Bernardi G (2000) Two classes of genes in plants. Genetics 154:1819–1825

Clay O, Caccio S, Zoubak S, Mouchiroud D, Bernardi G (1996) Human coding and noncoding DNA: compositional correlations. Mol Phylogenet Evol 5:2–12

Copenhaver GP, Nickel K, Kuromori T, Benito MI, Kaul S, Lin X, Bevan M, Murphy G, Harris B, Parnell LD, McCombie WR, Martienssen RA, Marra M, Preuss D (1999) Genetic definition and sequence analysis of *Arabidopsis* centromeres. Science 286:2468–2474

Eyre-Walker A (1999) Evidence of selection on silent site base composition in mammals: Potential implications for the evolution of isochores and junk DNA. Genetics 152:675–683

Eyre-Walker A, Hurst LD (2001) The evolution of isochores. Nat Rev Genet 2:549–555

Haupt W, Fischer TC, Winderl S, Fransz P, Torres-Ruiz RA (2001) The centromere l (CENl) region of *Arabidopsis thaliana:* Architecture and functional impact of chromatin. Plant J 27:285–296

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Li W (2001) Delineating relative homogeneous G + C domains in DNA sequences. Gene 276:57–72

Li W, Bernaola-Galvan P, Haghighi F, Grosse I (2002) Applications of recursive segmentation to the analysis of DNA sequences. Comput Chem 26:491–510

Macaya G, Thiery JP, Bernardi G (1976) An approach on the organization of eukaryotic genomes at a macromolecular level. J Mol Biol 108:237–254

Matassi G, Montero LM, Salinas J, Bernardi G (1989) The isochore organization and the compositional distribution of homologous coding sequences in the nuclear genome of plants. Nucleic Acids Res 17:5273–5290

McClintock B (1929) Chromosome morphology in *Zea mays*. Science 69:629

Meunier-Rotival M, Soriano P, Cuny G, Strauss F, Bernardi G (1982) Sequence organization and genomic distribution of the major family of interspersed repeats of mouse DNA. Proc Natl Acad Sci USA 79:355–359

Montero LM, Salinas J, Matassi Bernardi G (1990) Gene distribution and isochore organization in the nuclear genome of plants. Nucleic Acids Res 18:1859–1867

Nekrutenko A, Li WH (2000) Assessment of compositional heterogeneity within and between eukaryotic genomes. Genome Res 10:1986–1995

Oliver JL, Bernaola-Galvan P, Carpena P, Roman-Roldan R (2001) Isochore chromosome maps of eukaryotic genomes. Gene 276:47–56

Oliver JL, Carpena P, Roman-Roldan R, Mata-Balaguer T, Mejias-Romero A, Hackenberg M, Bernaola-Galvan P (2002) Isochore chromosome maps of the human genome. Gene 300:117–127

Pavlicek A, Paces J, Clay O, Bernardi G (2002) A compact view of isochores in the draft human genome sequence. FEBS Lett 511:165–169

Peshkin L, Gelfand MS (1999) Segmentation of yeast DNA using hidden Markov models. Bioinformatics 15:980–986

Rhoades MM (1978) In: Waraen BD (ed) Maize breeding and genetics. Wiley, New York, pp 641–672

Rhoades MM, Dempsey E (1973) Cytogenetic studies on a transmissible deficiency in chromosome 3 of maize. J Hered 64:13–18

Saccone S, De Sario A, Wiegant J, Raap AK, Della Valle G, Bernardi G (1993) Correlations between isochores and chromosomal bands in the human genome. Proc Natl Acad Sci USA 90:11929–11933

Salinas J, Matassi G, Montero LM, Bernardi G (1988) Compositional compartmentalization and compositional patterns in the nuclear genomes of plants. Nucleic Acids Res 16:4269–4285

Schroder AR, Shinn P, Chen H, Berry C, Ecker JR, Bushman F (2002) HIV-1 integration in the human genome favors active genes and local hotspots. Cell 110:521–529

Tenzen T, Yamagata T, Fukagawa T, Sugaya K, Ando A, Inoko H, Gojobori T, Fujiyama A, Okumura K, Ikemura T (1997) Precise switching of DNA replication timing in the GC content transition area in the human major histocompatibility complex. Mol Cell Biol 17:4043–4050

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, Antonarakis SE, Attwood J, Baertsch R, Bailey J, Barlow K, Beck S, Berry E, Birren B, Bloom T, Bork P, Botcherby M, Bray N, Brent MR, Brown DG, Brown SD, Bult C, Burton J, Butler J, Campbell RD, Carninci P, Cawley S, Chiaromonte F, Chinwalla AT, Church DM, Clamp M, Clee C, Collins FS, Cook LL, Copley RR, Coulson A, Couronne O, Cuff J, Curwen V, Cutts T, Daly M, David R, Davies J, Delehaunty KD, Deri J, Dermitzakis ET, Dewey C, Dickens NJ, Diekhans M, Dodge S, Dubchak I, Dunn DM, Eddy SR, Elnitski L, Emes RD, Eswara P, Eyras E, Felsenfeld A, Fewell GA, Flicek P, Foley K, Frankel WN, Fulton LA, Fulton RS, Furey TS, Gage D, Gibbs RA, Glusman G, Gnerre S, Goldman N, Goodstadt L, Grafham D, Graves TA, Green ED, Gregory S, Guigo R, Guyer M, Hardison RC, Haussler D, Hayashizaki Y, Hillier LW, Hinrichs A, Hlavina W, Holzer T, Hsu F, Hua A, Hubbard T, Hunt A, Jackson I, Jaffe DB, Johnson LS, Jones M, Jones TA, Joy A, Kamal M, Karlsson EK, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. Nature 420:520–562

Wright SI, Agrawal N, Bureau TE (2003) Effects of recombination rate and gene density on transposable element distributions in *Arabidopsis thaliana*. Genome Res 13:1897–1903

Zhang CT, Zhang R (1991) Analysis of distribution of bases in the coding sequences by a diagrammatic technique. Nucleic Acids Res 19:6313–6317

Zhang CT, Zhang R (2003) An isochore map of the human genome based on the Z curve method. Gene 317:127–135

Zhang CT, Wang J, Zhang R (200l) A novel method to calculate the G + C content of genomic DNA sequences. J Biomol Struct Dyn 19:333–341

Zhang R, Zhang CT (1994) Z curves, an intuitive tool for visualizing and analyzing the DNA sequences. J Biomol Struct Dyn 11:767–782

Zoubak S, Clay O, Bernardi G (1996) The gene distribution of the human genome. Gene 174:95–102